网络环境下人才知识结构的自动抽取方法*

刘庆祥[1] 张朋柱[1] 张晓燕[2] 刘景方[3]
[1](上海交通大学安泰经济与管理学院 上海 200030)
[2](上海工程技术大学管理学院 上海 201620)
[3](上海大学管理学院 上海 200444)

摘要: 【**目的**】构建人才知识结构的自动抽取方法。【**方法**】基于网络信息采集技术、网页分析以及文本分词、语义网相关技术, 构建基于网络环境的人才知识结构的自动抽取系统。【**结果**】实验验证了该系统的有用性, 系统识别课程的整体准确率在 95%以上, 对半结构化文件, 召回率在 95%以上; 对非结构化文件, 部分文件召回率低于 90%。【**局限**】课程识别的召回率受到词典库内容的制约。【**结论**】本方法能为人才知识结构研究提供有用的工具, 符合构建人才知识结构的基本要求。

关键词: 网络化创新外包 胜任力 知识结构 自动抽取
分类号: C931 G35

1 引 言

网络化创新服务外包, 指企业通过互联网利用外部人力资源完成创新任务的行为[1]。为实现对创新服务供需的迅速、精确匹配, 即人才能力和企业任务的有效匹配, 正确描述人才胜任力至关重要。1973 年美国著名心理学家 McClelland 首次提出人才胜任力的概念[2], 在此基础上 Mirable 对胜任力模型进一步总结, 提出了 KSAO 模型[3]。其中 K 表示知识, 指针对特定岗位和专业领域的要求具备的知识, 如岗位知识、专业知识, 本研究将主要围绕人才的专业知识结构的进行自动抽取。

当前人才知识结构信息获取方式仍然限于人工录入方式, 难以满足短时间获取大量数据的需求, 且成本相对较高, 发展一套可以自动化进行信息采集、分析、抽取的方法很有必要。

本研究的目的是构建一个可以从网络资源中自动抽取人才知识结构信息的方法, 利用丰富的网络资源进行人才的知识结构信息的采集、分析, 并自动抽取出人才知识库构建可利用的数据, 输出结果将考虑采用语义网、领域本体的相关技术和格式规范, 以便于对人才知识结构信息的灵活利用。

构建完备的人才知识库是完成人才能力和外包匹配的基础, 供给方只需提供人才的学校、专业、入学年份等信息, 系统即可从后台迅速地获取到人才的结构化的知识结构信息, 包括修读课程、课程内容等, 并可对不同的人才资源进行对比分析, 人才知识结构信息自动抽取系统的功能如图 1 所示:

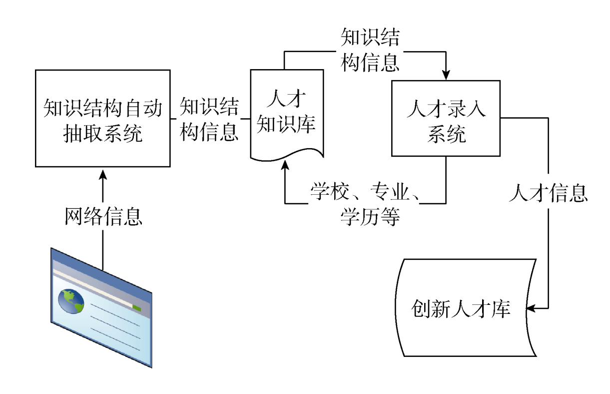


图 1 人才知识结构信息自动抽取系统示意图

通讯作者: 张晓燕, ORCID: 0000-0003-3189-8514, E-mail: zzlveofer@hotmail.com。

*本文系国家自然科学基金面上项目“网络化创新外包能力匹配”(项目编号: 71171131)、国家自然科学基金青年项目“在线群体创新中的图片支持研究”(项目编号: 71501124)和国家自然科学基金青年项目“众包环境下基于胜任力的供需双方匹配研究”(项目编号: 71301102)的研究成果之一。

2　人才知识结构自动抽取背景分析

当前我国大部分的高校在其官方网站、教务网站等渠道发布了其学院、专业、课程等信息，为本研究提供了丰富的信息来源，这将作为笔者主要的分析对象。

2.1　典型样本分析

在对比互联网来源的一些典型样本后，笔者总结出这些样本的关键特征：

(1) 信息量极大：据预估，合计开设课程的数量级在 10^7-10^9 之间，再考虑到人才的修读年份因素，人工整理难度极高。

(2) 变化频繁：在办学实践中，高校每年对院系、专业以及培养方案都会进行不同程度的调整，无疑为信息整理带来更多难题。

(3) 文件类型多样：各个高校在网站上发布的培养方案信息的形式多种多样，有 HTML、PDF、Word、Excel 等多种格式，系统的设计应当充分考虑多种文件格式的兼容。

(4) 半结构化/非结构化特征显著：各种文件内部的学院、专业、课程等信息的组织结构样式繁多，有以表格形式，如图 2 所示；也有纯文本形式的描述，如图 3 所示。

综合来看，信息的组织呈现出半结构化/非结构化特征，系统的设计应当针对不同结构特征的文件内容区分处理，以提高其处理信息的准确性。

必修：27 学分

课程号	课程名	周学时	学分	开课学期
00331751	微积分（一）	6	4	秋季（1）
00331770	线性代数与几何	5	4	秋季（1）
00331752	微积分（二）	6	4	春季（2）
00331860	高等微积分	4	3	秋季（3）
00331880	高等代数	3	3	秋季（3）
00330700	常微分方程	4	3	秋季（3）
00331900	概率与数理统计	3	3	秋季（5）
00330050	计算方法	5	3	春季（6）
		共	27	学分

图 2　信息组织示例 1：半结构化类型

本专业学生须按培养计划要求修读各类课程，总学分达到 220 学分，方可毕业。本专业所授学位为建筑学工学士。

建筑学专业毕业生修满以下专业类课程，在总学分满足要求的情况下，可取得建筑学工学士学位：

第 01 学期：设计概论、设计基础
第 02 学期：建筑概论、建筑设计基础
第 03 学期：建筑生成设计原理、建筑生成设计
第 04 学期：建筑设计原理、建筑设计
第 05 学期：公共建筑设计原理(1)-人文环境、公共建筑设计原理(2)-自然环境
公共建筑设计（人文环境与自然环境）
第 06 学期：公共建筑设计原理(3)-建筑群体、居住建筑设计原理
建筑群体设计与住区规划设计
第 07 学期：高层建筑设计原理、城市设计原理

图 3　信息组织示例 2：非结构化类型

(5) 内容差异性大：出于办学历史和特色原因，各高校的专业课程设置差异巨大。以管理信息系统专业为例，各高校的课程设置差异巨大[4]，且有的高校归属于计算机学院，有的归属于管理学院。

系统的设计应当充分考虑这种差异性，避免出现格式兼容上的问题。

2.2　系统设计要求与技术难点分析

(1) 尽可能最小化人工干预：信息量大且变化频繁的特征决定人工录入、整理的成本极高，因此信息的获取方式和知识结构的构建过程一定要满足自动化的要求，尽可能减少人工干预。

网络信息采集技术可满足上述需求，但是从头构建一个网络爬虫系统的工作量和复杂度极高，现有框架又难以满足个性化需求，应当充分考虑利用开源社区提供的成熟框架作为开发基础。

(2) 兼容性要求：文件类型的多样与内容的半结构化/非结构化特征要求系统对不同文件、结构类型兼容。

这对系统设计实现过程中的模块化与复用性提出较高的要求，应有前瞻的规划，采取扩展性良好的设计模式。

(3) 数据格式的扩展性要求：不同高校的专业课程设置内容的差异性对系统数据格式的设计提出要求，应当充分考虑知识结构构建过程中数据存储、转化的灵活性。

一般的关系型数据库所提供的存储方式较为单一，扩展性受限，应当参照当今 Web 数据广泛应用的 XML 等数据格式。

(4) 数据精度的要求：信息的精度包含信息获取、抽取的准确度，是保证人才知识库有用性的关键。

自动化的网络信息采集程序可迅速、批量地获取到大量网络资源，但这种方式难免带来有效信息纯度不高的问题，在系统分析设计的过程中，应当考虑集中数据源头、构建领域词典排除无用信息干扰、中间数据筛查等手段来抵消其不利影响。

2.3　输出结果要求

基于本研究的背景，为了实现最终资源后续利用的智能性与灵活性，输出结果要能够容易地被计算机读取并理解，具有一定的语义特征及本体实现。

本体构建的过程包含实体(Entity)、关系(object-Property)的分析，对所涉及的实体关系进行初步分析。

(1) 主要实体分析

研究涉及的实体有大学、院系、专业、课程等 4 种，如下：

①大学实体：一所大学由很多院系组成，大学本身具有校名、编号、介绍等属性；

②院系实体：一个院系是唯一地属于一个大学，院系下开设有很多专业。院系本身有院系概况的属性；

③专业实体：一个专业被开设在某个院系下(不是说一定属于某个院系)，开设很多课程，包括必修、选修课，其本身有概况、介绍等属性；

④课程实体：一个课程被开设在某个专业的培养计划中，也可能开设在其他专业中。课程本身有内容简介属性。

(2) 主要关系分析

实体间可能存在开设、被开设、隶属、拥有等关系。具体描述如下：

①opens 关系：开设关系，可应用于大学对院系、院系对专业、专业对课程，表明开设、拥有的关系，但不代表唯一拥有；

②is_opend_by 关系：被开设关系，是一种从属关系(例如一门课程可属于多个专业)，但不代表唯一从属关系，是 opens 关系的逆，应用于专业对院系，课程对专业的关系；

③associates_to 关系：唯一的从属关系，应用于院系对大学的关系(不同大学相同名称的院系视为不同的院系)；

④opens_as_required 和 opens_as_optional 关系：必修开设和选修开设关系，这是继承自 opens 的关系，应用于专业对课程的关系，区分这门课程在本专业中属于必修课还是选修课。

(3) 概念实体和属性关系图

概念实体和属性关系如图 4 所示：

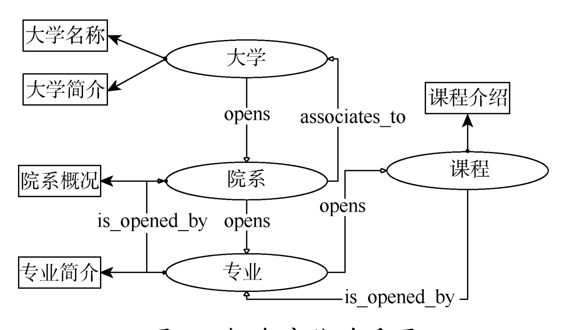


图 4　概念实体关系图

3　系统分析与设计

3.1　系统架构

从数据处理的角度出发，并结合前文要求，逐步分析以形成对系统的整体架构：

(1) 信息获取：采取网络爬虫程序从互联网上自动获取到大量人才知识结构信息原始数据。

(2) 初步转化：采用文本解析程序对类型各异的原始数据进行初步转化，得到对应的文本数据。

(3) 知识结构抽取：从不同结构类型的文本数据中进行学院、专业、课程实体识别，构建出相应的内存数据。

(4) 数据转储：为了方便数据的后续利用，将内存数据持久化为扩展性强的中间数据。

(5) 语义数据构建：以中间数据为基础，采取本体构建程序得到具有语义特征的数据，这些数据可补充词典库内容，提升系统对信息利用的完善程度。

综上，数据经历了原始数据、文本数据、内存数据、中间数据、语义数据的转化流程，如图 5 所示：

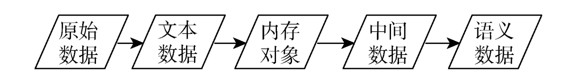


图 5　数据格式转化

由此得出系统的整体架构如图 6 所示：

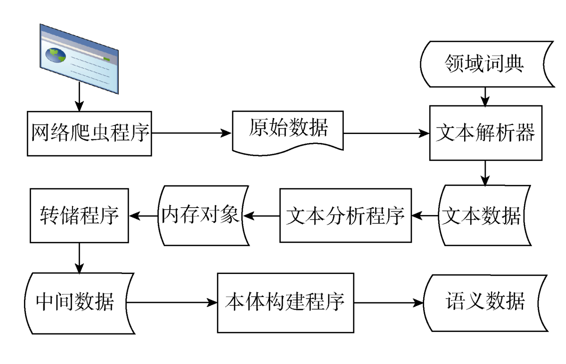


图 6　系统设计图

3.2　网络爬虫程序

网络资源的遍历一般有广度优先算法和深度优先算法，笔者采用广度优先算法，一般来说越有价值的专业课程信息距离数据源头越近，也越有价值。

除了选择适当的搜索算法外，还应当遵循适当的 URL 和文件过滤规则，以提高系统的抓取效率：

(1) URL 过滤：因为在本文中，抓取信息来源于国内高校官方网站，对于不含“edu”的网页将被过滤而不进行抓取；

(2) 文件过滤：对于 HTML 网页，程序将设定一系列的关键词集合，读取其文本信息与其进行匹配，不含任何关键词集合的 HTML 网页将不进行抓取，关键词集合包含“专业”、“学院”、“课程”等。

3.3　文件解析程序

鉴于原始数据的多样性, HTML、PDF、Office 类型(Word、Excel)等文件难以被直接利用, 因此有必要通过一定的技术手段进行解析, 将各类原始文件统一转化为容易利用的文本数据。文件解析器程序分别调用不同的文件解析接口, 根据文件类型进行区别处理, 流程如图 7 所示:

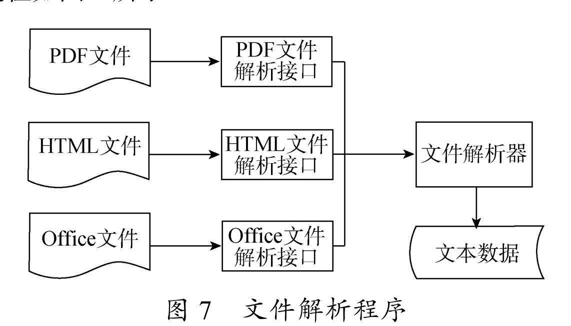


图 7　文件解析程序

3.4　文本分析与转储程序

文本分析读取文件解析程序返回的文本数据, 对文本内容进行分析后得到内存对象。文本分析流程如图 8 所示:

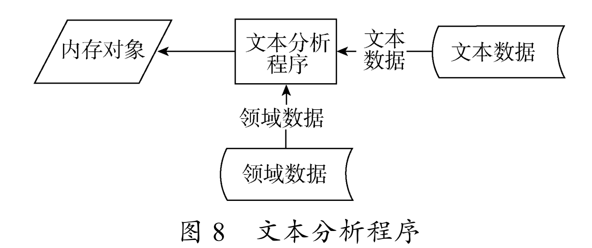


图 8　文本分析程序

对半结构化的文本内容, 不需要读取领域词典即可得到层次分明的专业课程信息。程序读取文本数据中的表格数据, 逐次读取表格内容, 并根据其对应表头内容将专业、课程信息映射到内存对象中, 例如从培养方案表格中将课程名称、学分、学时的情况映射到一个 Course 对象, 对应其 name、credit、period 属性(可以为空)。

而对于非结构化文本, 内容相对散乱无序, 分析的过程更为复杂, 要依据一定的领域词典匹配。程序将读取所有文本内容信息, 依据系统的领域词典中的专业、课程词汇进行分词, 对于课程实体, 将在实体所属语句内搜索课程属性关键词(学分、学时、介绍等), 定位属性内容, 映射到 Course 对象。将专业课程的内存对象数据转储为中间数据。

3.5　本体构建程序

利用返回的中间数据, 将它们按照最终对信息结构的需求构建出语义本体数据, 以增强信息被计算机读取、理解的强度, 适应可能出现的信息的灵活利用要求。本体构建程序如图 9 所示:

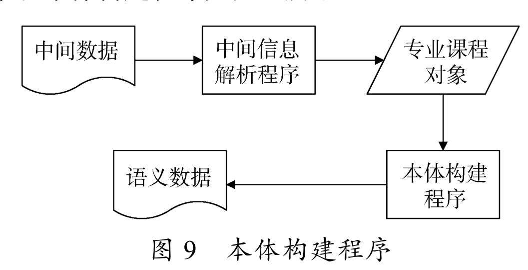


图 9　本体构建程序

4　数据结构设计

4.1　爬虫数据库设计

爬虫数据库的备选的方案包括关系型数据库例如 Oracle、SQL Server 等, 文件型数据结构 XML、RDF 等, 以及嵌入式数据库 Berkeley DB 等。

考虑访问效率、安全性、海量数据负载等要求, 笔者采取 Berkeley DB 作为爬虫数据库。

Berkeley DB 采取关键字/数据(Key/Value)的形式进行数据库管理, 通过相关 API, 提供关键字即可获取到对应数据, 访问效率很高, 其底层可以理解为存放大量数据的 HashMap, 访问复杂度只有 O(1), 性能要明显优于关系型、文件型数据库。Berkeley DB 的数据结构如下:

(1) BdbFrontier: 这个类是 Heritrix 中使用 Berkeley DB 结构的链接制造工厂, 用来验证某个正等待进入队列的对象是否已被抓取过。

(2) BdbMultipleWorkQueues: 一组链接对象的队列, 不同的队列具有不同的 Key 值, 即 Key 和链接队列形成一个个"Key/Value"对, 作为 Berkeley DB 中的一条记录。如表 1 所示:

表 1　BdbMulipleWorkQueues 结构示意

键	值
Key1	Queue1{URI1, URI2, URI3, …}
Key2	{…}
Key3	{…}
……	……

(3) BdbWorkQueue: 基于 Berkeley DB 的链接队列, 创建每个 BdbWorkQueue 都会赋予一个键值。

(4) BdbUriUniqFilter: 过滤器, 被 BdbFrontier 调用, 内部包含被抓取过链接的 Berkeley DB 数据库[5]。

4.2 中间数据格式设计

可参考格式有 XML、SQL Server、MySQL 等。考虑向语义数据转化的难度以及数据迁移的方便后，决定选取 XML 格式。XML 文件被广泛认为是语义网实现的基础层，语法标准统一，可扩展性明显优于关系型数据库。笔者给出 XML 文件规范 format.xsd 结构如图 10 所示:

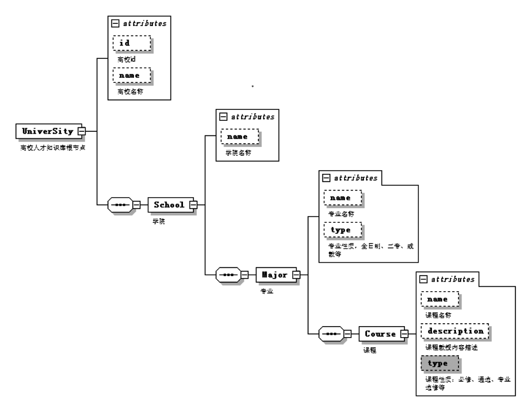


图 10 中间数据格式

4.3 语义数据格式设计

语义数据格式设计的备选方案有 OWL、XML、RDF 等，考虑表达能力的需求，本研究优先考虑采用 OWL 的格式。OWL(Web 本体语言)是和语义 Web 相关的 W3C 推荐标准栈的一部分，使用基于 XML 的 RDF 语法[6]，表达能力要远强于 XML 和 RDF。语义数据格式设计参照图 4。

5 系统实现与测试

5.1 编程语言与开发环境

本系统采用 Java 语言为程序开发语言，与 C、C++相比，Java 彻底面向对象，设计模式运用方便，且由于跨平台特性，可直接调用的成熟开源框架很多。系统采用 Eclipse 作为开发环境，与其他开发工具如 JBuilder、IDEA 相比，Eclipse 具备开放、自由、可扩展插件众多等优势，能满足快速开发的需求。

5.2 网络爬虫程序与运行效果

考虑到系统复杂度高、个性化需求多等难点，笔者决定选取开源的爬虫框架作为开发基础，有 Scarpy、Cola、Heritrix、Beautiful Soup 等可供选取，在充分考虑编程语言、界面友好性、扩展性因素后，最终决定采用 Heritrix 框架作为基本框架。

Heritrix 是一个始于 2003 年的开源、可扩展的网络爬虫项目，基于 Java 平台开发[7]，与 Scrapy、Cola 相比，其配置功能更为强大，具有更好的扩展性，而且它可以通过 Web 界面操作，友好性更强，其基本框架如图 11 所示:

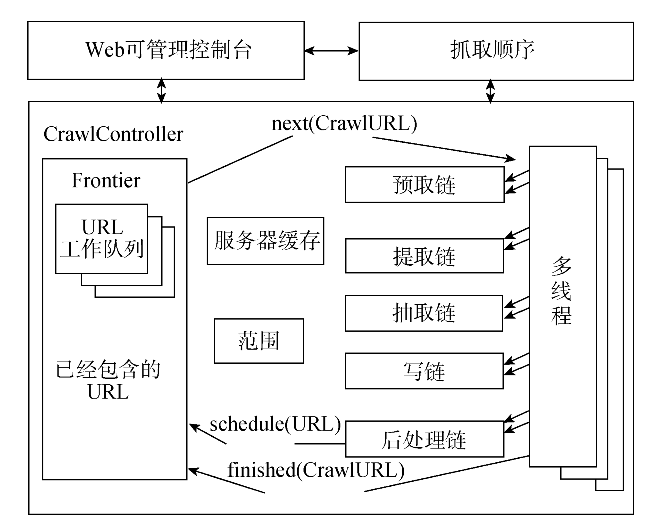


图 11 Heritrix 基本框架

其中 CrawlController 类(控制器)协调各模块的运行，是本框架的核心，CrawlController 作为爬虫系统的中枢神经，决定爬虫进程的开始、结束。其核心有三大部件：范围部件、边界(Fontier)部件、处理器链。

(1) 范围部件决定下一个入列的 URL 是什么，可自定义干涉;

(2) 边界部件进行边界条件的验证，对未访问队列中的 URL 进行验证，上文所述的 URL 规则即在此部分设定;

(3) 处理器链中是正在同时处理的 URL 队列，处理结果会传递给边界条件。

爬虫任务的创建界面如图 12 所示。这是一个以上海交通大学本科教学信息服务网为源头的任务，数据源头如 Seeds 中所示，再对 Modules、Setting 模块的参数进行设定，任务即创建成功。

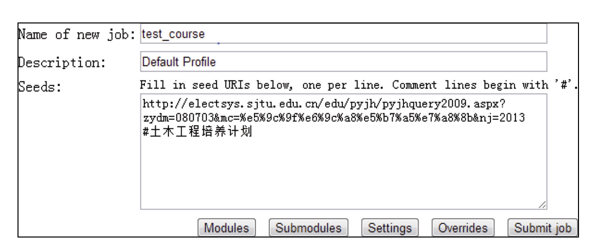
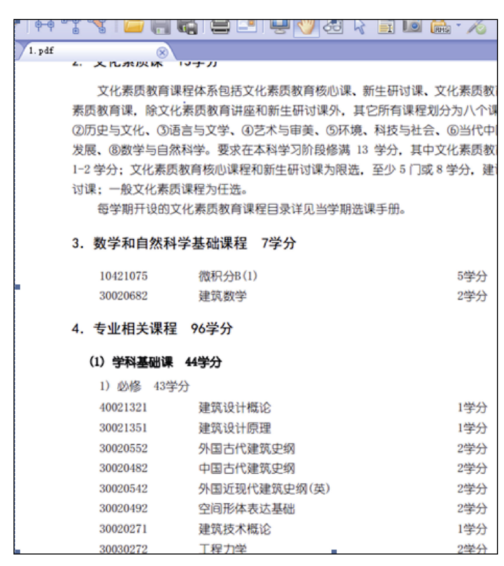


图 12 爬虫任务创建

5.3　文件解析程序与运行效果

文件解析器通过分别调用不同的接口，将文件映射到内存再从中读取出要利用的字符串，输出文本文件。以 PDF 文件的解析为例，阐述解析程序的搭建：采用 PDFBox API 进行开发。它采用面向对象的方式获取 PDF 文档，不同于文本格式的文件流，将一个 PDF 文件视为一系列基本对象的组合，包含数组、数字、字符串、词典等结构，非常适合本研究的开发框架。

待转化的 PDF 测试文件，如图 13 所示。这是某所高校建筑学专业的培养计划文件。转化所得到的文本文档如图 14 所示。



1.pdf

文化素质教育课程体系包括文化素质教育核心课、新生研讨课、文化素质教
素质教育课，除文化素质教育讲座和新生研讨课外，其它所有课程划分为八个课
②历史与文化、③语言与文学、④艺术与审美、⑤环境、科技与社会、⑥当代中
发展、⑧数学与自然科学。要求在本科学习阶段修满 13 学分，其中文化素质教
1-2 学分；文化素质教育核心课程和新生研讨课为限选，至少 5 门或 8 学分，建
讨课；一般文化素质课程为任选。
每学期开设的文化素质教育课程目录详见当学期选课手册。

3. 数学和自然科学基础课程　7学分

10421075	微积分B(1)	5学分
30020682	建筑数学	2学分

4. 专业相关课程　96学分

(1) 学科基础课　44学分

1) 必修　43学分

40021321	建筑设计概论	1学分
30021351	建筑设计原理	1学分
30020552	外国古代建筑史纲	2学分
30020482	中国古代建筑史纲	2学分
30020542	外国近现代建筑史纲(英)	2学分
30020492	空间形体表达基础	2学分
30020271	建筑技术概论	1学分
30030272	工程力学	2学分

图 13　处理前 PDF 文档

1.txt - 记事本
文件(F)　编辑(E)　格式(O)　查看(V)　帮助(H)
清华大学本科培养方案
1
建筑系
建筑学专业本科培养方案
一、培养目标
本专业是一门综合性较强的学科，涉及理、工、文、艺诸领域，具有科
合的特点。本专业立足中国、面向世界，培养符合国家建设需求，了解
具备成为专业帅才潜力的高素质建筑人才。
二、基本要求
建筑学专业毕业生应达到以下知识、能力和素质的要求：
具备高尚的道德素质、丰富的人文素质、健康的身心素质。
具备建筑理论、建筑历史和建筑技术以及城乡规划、风景园林等方面的
筑相关领域知识。
具有严谨的科学精神和较强的创新思维，具备整合建筑形象思维和逻辑
设计和实践能力。
具有社会责任、职业道德、团队精神、跨专业协调能力和综合管理与领
三、学制与学位授予
本专业设置两种学制：
1.本科学制五年，按照学分制管理机制，实行弹性学习年限。授予学位：
2.本科学制四年，按照学分制管理机制，实行弹性学习年限。授予学位：
四、基本学分学时
四年制本科培养总学分 175，其中春、秋季学期课程总学分 142；集中
实践环节 18学分，综合论文训练 15学分。
五年制本科培养总学分 200，其中课程学分 156；集中实践环节由两部
综合论文训练 15学分。
五、专业核心课程　9门，42学分
建筑设计(1)(6学分)、建筑设计(2)(6学分)、建筑设计(3)(6学分)、建
筑设计(5)(6学分)、设计专题(1)(3学分)、设计专题(2)(3学分)、设计
(4)(3学分)。
六、课程设置与学分分布
四年制本科课程设置与学分分布

图 14　转化后文本文档

5.4　文本分析与转储程序实现与运行效果

将获取的文本信息转化成 XML 文件的几个关键的方法描述如下：

txtToXml(String in, String out); //主方法，接收文本文件的路径和输出 XML 文件的路径，完成转化。

coursesToXml(List<Course> courses, String out); //txtToXml 调用，接收 XML 文件的路径，将课程对象列表持久化到 XML 文件。

parseOneCourse(String line); //接收文本文件的一行内容，返回一个 Course 对象或者 Null。中间数据生成程序调用 Dom4j 接口来实现 XML 文件构造。

5.5　语义数据构建和测试效果

本体构造工具种类繁多，包括商业产品、高校与研究机构的课题成果[8]，有 OntoEdit、WebOnto、Protégé、WebODE 等。本文选取 Protégé 作为构建工具。Protégé 是斯坦福大学开发的基于 Java 语言的本体编辑、知识获取软件[9]，拥有很多优秀设计的插件，其扩展性、友好性明显强于其他工具，是当前使用最广泛的本体编辑器。

本体构建工具搭建完成后，选取测试数据进行相似度计算，如下为测试数据描述：

S 大学下有计算机科学与技术系和软件工程学院，其中计算机科学与技术系下有计算机科学与技术专业，必修课程有 C++程序设计，数据结构，软件工程概论，算法与复杂性，选修课程有海量数据处理……软件工程学院下有软件工程专业，C++程序设计，数据结构，软件工程概论，IT 服务管理，选修课程有中间件技术……

测试结果如表 2 所示：

表 2　相似度计算结果

距离	相似度
Hier Distantce	0.8000
Attr Distance	0.1670
Distance	0.4583

6　实验效果及分析

为验证自动抽取方法的有用性，组织了一次知识结构信息抽取的实验。

6.1　实验内容

以 50 所高校、超过 2 000 个专业的培养方案文件为原始数据，文件大小为 834MB。

根据文件内容的结构分为两组，即半结构化组与非结构化组，在本研究中分别指表格类型和松散文本类型的文件内容，分组后得到 37 组半结构化组文件

和 13 组非结构化组文件。从两组文件中分别随机选取 10 所高校，每所随机选取 10 个专业，得到 200 个专业的培养方案，人工分割成 200 个文件。由于文件格式不一，单个文件大小在 10-200K 之间，文件大小总计 23.1MB。以这些文件为原始数据，进行处理，得到课程识别的结果，再由两名实验人员人工核对，计算准确率和召回率。

为方便实验人员比对，将所得结果转化为 Excel 格式，结果如图 15 所示：

	A	B	C
1	物联网工程	电子信息工程	电子科学与技术（微电子学与固体电子学）
2	思想道德修养与法律基础	思想道德修养与法律基础	思想道德修养与法律基础
3	中国近现代史纲要	中国近现代史纲要	中国近现代史纲要
4	马克思主义基本原理	马克思主义基本原理	马克思主义基本原理
5	毛泽东思想和中国特色社会主义理论体系概	毛泽东思想和中国特色社会主义理论体系概	毛泽东思想和中国特色社会主义理论体系概
6	形势与政策教育	形势与政策教育	形势与政策教育
7	英语读写译1A - D	英语读写译1A - D	英语读写译1A - D
8	英语口语1A - D	英语口语1A - D	英语口语1A - D
9	英语听力1A - B	英语听力1A - B	英语听力1A - B
10	英语读写译2A - D	英语读写译2A - D	英语读写译2A - D
11	英语口语2A - D	英语口语2A - D	英语口语2A - D
12	英语听力2A - B	英语听力2A - B	英语听力2A - B
13	翻译	翻译	翻译
14	商务英语	商务英语	商务英语
15	英美文化	英美文化	英美文化
16	影视欣赏	影视欣赏	影视欣赏
17	西方文化	西方文化	西方文化
18	英美文学简史	英美文学简史	英美文学简史
19	法制安全教育	法制安全教育	法制安全教育
20	管理概论	管理概论	管理概论
21	职业生涯规划	职业生涯规划	职业生涯规划

图 15　实验生成文件

6.2　评价指标

对实验结果的评价采用自然语言处理中常用的准确率(Precision)、召回率(Recall)指标。准确率是衡量信息检索结果的质量，即查准率；召回率用来衡量信息检索结果的查全率。笔者定义评价指标，如表 3 所示：

表 3　实验评价指标

名称	缩写	含义解释
单篇准确率	SP	单个文件中课程识别正确数/课程识别数
单篇召回率	SR	单个文件中课程识别数/课程总数
平均准确率	AP	全部文件中课程识别正确数/课程识别数
平均召回率	AR	全部文件中课程识别数/课程总数
准确率标准差	Std.P	单篇准确率的标准差
召回率标准差	Std.R	单篇召回率的标准差

6.3　实验分析

实验结果如表 4 所示：

表 4　实验结果统计

指标	半结构化组	非结构化组
N	100	100
AP(%)	97.96	99.46
Std.P	1.102×0.01	0.300×0.01
AR(%)	99.51	89.39
Std.R	0.291×0.01	5.411×0.01

实验结果表明，本系统对于输入文件的平均准确率较高，在 95%以上。对于半结构化组的培养方案文件，系统的召回率较高，在 99%以上，但是对于非结构化类型的文件，系统的平均召回率较低，在 90%以下，实验中某些文件的课程识别率在 80%左右，这主要是由于对于处理非结构化文件，系统比较依赖词典库的完善度，对于词典库中缺乏的课程词汇，往往难以识别，例如“程序设计方法与思想”课程，在某些学校名称是“计算思维”，如果词典库中没有该词汇，将难以判别。但是，随着系统处理结构化/半结构化文件的数量增多，词典库内容会进一步扩充，对非结构化文件的识别召回率也会有所提高。

7　结　语

本研究详细分析了网络化创新外包中的人才知识结构抽取的背景，并设计实现了人才知识结构信息的自动抽取系统，得到的输出结果可以为构建人才知识库提供支撑。依赖本研究成果，只需人才的少量基本信息即可迅速获取到他们在高校所修读的课程及其描述，并可依据最终的语义数据给出相似度的分析结果，在今后的数据更新中，也不必再耗费大量的人力成本。

本文成果将会在未来的人才知识库、创新任务匹配的研究中发挥作用，庞大的专业课程数据将成为人才知识库构建的强大后备，最终的语义数据也会逐步提升系统课程识别的召回率。在实践中，知识结构不仅仅包含人才在高校接受教育所掌握的知识，也包含人才在后来的培训、工作之中所掌握的技能、职业的专业知识，这些知识采取什么样的形式构建、存储都可以参照本文所示的构建方式，后续研究也将围绕这些范畴和特定专业领域展开。

参考文献：

[1] 李小卯，张建军. 基于 Internet 的资源外包与企业创新[J]. 中国软科学，2003(1): 93-99. (Li Xiaomao, Zhang Jianjun. Internet-based Outsourcing and Enterprise Innovation [J]. China Soft Science, 2003(1): 93-99.)

[2] McClelland D C. Testing for Competence Rather than for Intelligence [J]. American Psychologist, 1973, 28(1): 1-14.

[3] Mirable R J. Everything You Wanted to Know About

Competency Modeling [J]. Training and Development, 1997, 51(8): 73-77.

[4] 何永刚, 黄丽华. 信息管理与信息系统专业课程体系研究综述[J]. 情报杂志, 2007, 26(8): 128-131. (He Yonggang, Huang Lihua. Research Reviews on the Curriculums for the Information Systems [J]. Journal of Information, 2007, 26(8): 128-131.)

[5] 邱哲, 符滔滔. 开发自己的搜索引擎: Lucene 2.0+Heritrix [M]. 北京: 人民邮电出版社, 2007. (Qiu Zhe, Fu Taotao. Develop Your Own Search Engine: Lucene 2.0+Heritrix[M]. Beijing: Posts & Telecom Press, 2007.)

[6] 高志强. 语义 Web 原理及应用[M]. 北京: 机械工业出版社, 2009. (Gao Zhiqiang. The Principle and Application of Semantic Web [M]. Beijing: China Machine Press, 2009.)

[7] 罗刚, 王振东. 自己动手写网络爬虫[M]. 北京: 清华大学出版社, 2010. (Luo Gang, Wang Zhendong. Build a Web Crawler by Yourself [M]. Beijing: Tsinghua University Press, 2010.)

[8] 徐国虎, 许芳. 本体构建工具的分析与比较[J]. 图书情报工作, 2006, 50(1): 44-48. (Xu Guohu, Xu Fang. A Comparative Study of Ontology-building Tool [J]. Library and Information Service, 2006, 50(1): 44-48)

[9] 洪娜, 张智雄. Protégé 在科研本体构建与推理中的实践研究[J]. 现代图书情报技术, 2009(7-8): 1-5. (Hong Na, Zhang Zhixiong. Practice of Creating and Reasoning Science Ontology by Protégé [J]. New Technology of Library and Information Service, 2009(7-8): 1-5.)

作者贡献声明:

刘庆祥: 提出论文思路, 系统开发, 论文起草及最终版本修订;
张朋柱: 研究思路提出, 研究方案设计;
张晓燕: 提出实验设计方案, 采集、清洗和分析数据;
刘景方: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: kentliu1990@163.com。
[1] 刘庆祥. knlgbuild.xls. 人才知识结构自动抽取实验.
[2] 刘庆祥. crowdsourcing.zip. 人才知识结构自动抽取系统.

收稿日期: 2015-12-09
收修改稿日期: 2016-01-08

Automatically Extracting Talents' Knowledge Structure Online

Liu Qingxiang[1]　Zhang Pengzhu[1]　Zhang Xiaoyan[2]　Liu Jingfang[3]
[1](Antai College of Economics and Management, Shanghai Jiaotong University, Shanghai 200030, China)
[2] (School of Management, Shanghai University of Engineering Science, Shanghai 201620, China)
[3](School of Management, Shanghai University, Shanghai 200444, China)

Abstract: [**Objective**] To extract talents' knowledge structure automatically. [**Methods**] We built an online knowledge structure extraction system based on Web information retrieval, webpage analysis, word segmentation and semantic Web technologies. [**Results**] We examined the usability of the new system. For course recognition, the overall precision rate was more than 95%. For semi-structured files, the recall rate was above 95%. For some non-structured files, the reacall rate was below 90%. [**Limitations**] The recall rate of course recognition was restricted by the content of the dictionary. [**Conclusions**] The proposed method meets the requirements of constructing talents' knowledge structure and is a useful tool for related research.

Keywords: Web-based outsourcing　Competency　Knowledge structure　Automatic extraction